NicenessMovies

Progetto Fondamenti di Intelligenza Artificiale

Matricola: 0512113708

Giulio De Pascale

26/01/2024

https://github.com/Gdp-22/LetterboxNiceness

Sommario

1 Introduzione

1.1 L'idea

L'idea prende forma dall'utilizzo di 'Letterboxd', un social network che consente agli utenti di cercare film e visualizzarne le rispettive caratteristiche. La funzione centrale di questo social network, tuttavia, risiede nella possibilità di esprimere valutazioni per ciascun film.

1.2 La natura del problema

Nel contesto di piattaforme di streaming per film, si possono individuare numerosi modelli e sistemi avanzati di raccomandazione concepiti per suggerire i migliori contenuti da guardare, basandosi sulle attività dell'utente all'interno di una piattaforma. Tuttavia, quando un utente è indeciso su un film specifico, sorge la problematica di non sapere in anticipo se il film scelto sarà di suo gradimento. I sistemi più noti, forniscono solamente una media della valutazione globale del film. L'utente potrebbe trovarsi nella situazione di iniziare a guardare un film senza avere certezze sul suo apprezzamento personale, scoprendo magari, dopo una ventina di minuti, che il film non è di suo gradimento.

1.3 Sistema proposto

Il sistema proposto consentirà all'utente di ottenere una metrica più precisa sulla probabilità di gradimento di un determinato film. Ciò sarà reso possibile grazie a un indice di gradevolezza associato al film cercato, il quale terrà conto del profilo storico dell'utente in termini di valutazioni assegnate a film precedentemente visualizzati. In questo modo, l'utente potrà beneficiare di un'indicazione più realistica riguardo alle sue preferenze, migliorando così la sua esperienza nella selezione dei contenuti .

2 Specifica PEAS

PEAS	
Performanace	La misura di performance dell'agente è la sua capacità di avvicinarsi quanto più possibile ad una situazione ideale nella quale all'utente per ogni specifico film viene mostrato il giusto indice di gradevolezza.
Enviroment	L'ambiente in cui opera l'agente è lo spazio dei film visti con le loro valutazioni valutazioni unito a tutti i possibili film da vedere e le loro caratteristiche
Actuators	Predizione dell'indice di gradevolezza del film
Sensors	Ricezione delle valutazioni utente, informazioni del film

2.1 Proprietà dell'ambiente

- **Completamente osservabile**, Si ha accesso a tutte le caratteristiche di un film e alle valutazioni.
- **Deterministico**, Le predizioni non hanno una componente di casualità
- **Episodico**, Le predizioni passate non influenzano quelle future
- **Statico**, I film e le valutazioni non cambiano nel tempo.
- **Discreto**, Le variabili assumono valori in un intervallo limitato.
- **Singolo agente**.

3 CRISP-DM

Per l'implementazione è stato utilizzato il modello progettuale CRISP-DM

3.1 Bussiness Understanding

Il principale criterio di successo del progetto è di avere un modello di AI che riesca a predire la gradevolezza di un determinato film sulla base delle valutazioni dell'utente su altri film.

3.2 Data Understanding

I dati sono stati ottenuti da due diverse fonti: il primo dataset, denominato 'ratingsMe' e rappresentato nella Figura 1, è stato generato tramite il social network Letterboxd. Questo dataset è stato creato sfruttando la funzione di esportazione dati messa a disposizione direttamente dal social, accessibile attraverso il link esporta-dati. Al suo interno, sono contenute tutte le valutazioni relative ai rispettivi film, fornite dall'utente.
Il secondo dataset, denominato 'IMDbMovies' e presentato nella Figura 2, è stato scaricato da Kaggle. Questo dataset comprende le caratteristiche di un ampio numero di film.

index	Name	Year	Rating
0	8½	1963	5.0
1	Black Widow	2021	1.5
2	The Seventh Seal	1957	5.0
3	Death Proof	2007	4.0
4	Knives Out	2019	3.0
5	Inception	2010	3.0
6	The Dark Knight	2008	3.5
7	Interstellar	2014	3.0
8	Dunkirk	2017	3.5
9	Memento	2000	4.0
10	Tenet	2020	2.5
11	Following	1998	2.5

Figure 1: ratingsMe

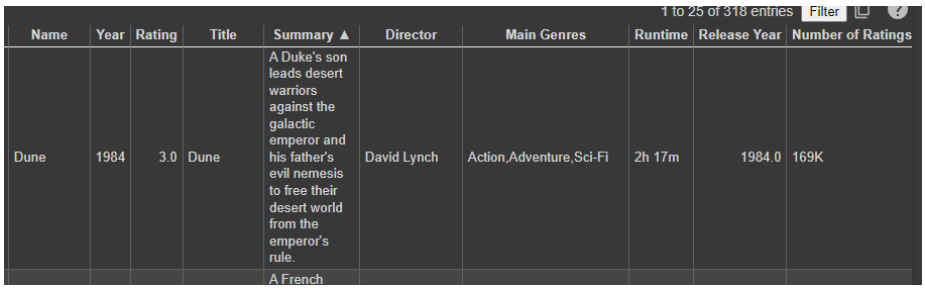
index	Title	Summary	Director	Main Genres	Runtime	Release Year	Rating	Number of Ratings
0	Napoleon	An epic that details the checkered rise and fall of French Emperor Napoleon Bonaparte and his relentless journey to power through the prism of his addictive, volatile relationship with his wife, Josephine.	Ridley Scott	Action,Adventure,Biography	2h 38m	2023.0	6.7/10	38K
1	The Hunger Games: The Ballad of Songbirds &	Coriolanus Snow mentors and develops feelings for the female District 12 tribute during the	Francis Lawrence	Action,Adventure,Drama	2h 37m	2023.0	7.2/10	37K

Figure 2: IMDBMovies

I due set di dati vengono combinati in un unico dataset, come illustrato nella Figura 3, utilizzando i campi "titolo" e "anno" come chiavi di unione. Questa operazione consente di consolidare le informazioni relative ai film sottoposti a valutazione, garantendo che i dati pertinenti da entrambi i dataset siano inclusi in modo coerente.

3.2.1 Semantica degli attributi

- Title: Titolo del film
- Director: Regista del film
- Main Genres: Generi principali del film
- Runtime: Durata del film
- Year: Anno di uscita del film
- Rating: Valutazione dell'utente per il film
- AvgRating: Media delle valutazioni degli utenti per il film
- Number of Ratings: Numero delle valutazioni degli utenti per il film



1 to 25 of 318 entries Filter

Name	Year	Rating	Title	Summary ▲	Director	Main Genres	Runtime	Release Year	Number of Ratings
Dune	1984	3.0	Dune	A Duke's son leads desert warriors against the galactic emperor and his father's evil nemesis to free their desert world from the emperor's rule.	David Lynch	Action,Adventure,Sci-Fi	2h 17m	1984.0	169K
				A French					

Figure 3: Unione Dataset

3.3 Data Preparation

3.3.1 Data cleaning

Per quanto riguarda le operazioni di data cleaning, le colonne considerate superflue per la predizione sono state rimosse; le righe con valori nulli sono veramente poche, quindi si è deciso di eliminarle. I dati con un formato complesso, inoltre, sono stati formattati. Le valutazioni sono espresse in termini di numeri interi, su scala da 1 a 10, La durata del film è espressa in minuti, Il numero di valutazioni è espresso in interi.

Year	Rating	Director	Main Genres ▲	Runtime	Number of Ratings	AvgRating
1989	8	Tim Burton	Action,Adventure	126	398000	7
2023	7	Ridley Scott	Action,Adventure,Biography	158	38000	6
2014	7	Seth Rogen,Evan Goldberg	Action,Adventure,Comedy	112	348000	6
2021	7	James Gunn	Action,Adventure,Comedy	132	393000	7
1980	8	John Landis	Action,Adventure,Comedy	133	210000	7
2014	8	James Gunn	Action,Adventure,Comedy	121	1300000	8
2017	8	James Gunn	Action,Adventure,Comedy	136	742000	7
2017	4	Taika Waititi	Action,Adventure,Comedy	130	800000	7
2023	6	John Francis Daley,Jonathan Goldstein	Action,Adventure,Comedy	134	202000	7
2023	8	James Gunn	Action,Adventure,Comedy	150	347000	7
2000	9	Ridley Scott	Action,Adventure,Drama	155	1600000	8
2001	8	Peter Jackson	Action,Adventure,Drama	178	2000000	8
2021	8	Denis Villeneuve	Action,Adventure,Drama	155	729000	8

Figure 4: Dataset after data cleaning

3.3.2 Categorizzazione delle feature

Viste le distribuzioni molto ampie dei valori si decide di categorizzare le feature:

- Year: 'Recente', 'Intermedio', 'Storico'
- Rating: 'Discreto', 'Buono', 'Ottimo'
- Runtime: 'Lunga', 'Corta', 'Media'
- Number of ratings: 'Sconosciuto', 'Conosciuto', 'Popolare'
- AvgRating: 'Bassa', 'Media', 'Alta'

I valori sono stati associati alle menzionate etichette non in base alla distanza nello spazio di valori, bensì con l'obiettivo di bilanciare il dataset per ciascuna etichetta di una determinata caratteristica. Tale assegnazione è stata effettuata in conformità con la distribuzione dei valori, come evidenziato nella Figura 5.
Questo approccio ha reso superflua la fase di data balancing.

```
df.describe()
```

1 to 8 of 8 entries Filter

index	Year	Rating	Runtime	Number of Ratings	AvgRating
count	316.0	316.0	316.0	316.0	316.0
mean	2000.3449367088608	7.800632911392405	121.56962025316456	437480.7088607595	7.069620253164557
std	20.940838956127276	1.8792432396331913	26.715400014796195	490644.6510861193	0.825137515100727
min	1920.0	1.0	64.0	504.0	3.0
25%	1987.75	7.0	101.0	100500.0	7.0
50%	2006.0	8.0	117.0	250000.0	7.0
75%	2017.0	9.0	136.0	643000.0	8.0
max	2023.0	10.0	229.0	2800000.0	9.0

Figure 5: Distribution of values

Year	Rating	Director	Main Genres ▲	Runtime	Number of Ratings	AvgRating
Storico	Buono	Tim Burton	Action,Adventure	Lunga	Conosciuto	Media
Recente	Discreto	Ridley Scott	Action,Adventure,Biography	Lunga	Sconosciuto	Bassa
Recente	Discreto	Taika Waititi	Action,Adventure,Comedy	Lunga	Popolare	Media
Intermedio	Discreto	Seth Rogen,Evan Goldberg	Action,Adventure,Comedy	Media	Conosciuto	Bassa
Storico	Buono	John Landis	Action,Adventure,Comedy	Lunga	Conosciuto	Media
Recente	Discreto	John Francis Daley,Jonathan Goldstein	Action,Adventure,Comedy	Lunga	Conosciuto	Media
Recente	Discreto	James Gunn	Action,Adventure,Comedy	Lunga	Conosciuto	Media

Figure 6: Dataset after categorization

3.3.3 Feature Engineering

In fase di Feature Engineering sono state aggiunte diverse feature:

- Notorietà, rappresnta la cosiderazione dell'utente verso il regista del film; può essere 'Preferito', 'Apprezzato', 'Noto', 'Sporadico'
- Apprezzato, rappresenta la considerazione globale degli utenti per il film; può essere 'Si', 'No',
- Main Genre, rappresenta il principale genere del film, la feature è estratta da 'Main Genres'.

Year	Rating	Director	Main Genres	Runtime	Number of Ratings	AvgRating	Notorieta	Apprezzato	Main Genre
Storico	Ottimo	Federico Fellini	Drama	Lunga	Sconosciuto	Alta	Noto	No	Drama
Recente	Discreto	Cate Shortland	Action,Adventure,Sci-Fi	Lunga	Conosciuto	Bassa	Sporadico	No	Action
Storico	Ottimo	Ingmar Bergman	Drama,Fantasy	Corta	Conosciuto	Alta	Sporadico	Si	Drama
Intermedio	Buono	Quentin Tarantino	Action,Thriller	Lunga	Conosciuto	Media	Preferito	No	Action
Recente	Discreto	Rian Johnson	Comedy,Crime,Drama	Lunga	Popolare	Media	Sporadico	No	Comedy
Intermedio	Discreto	Christopher Nolan	Action,Adventure,Sci-Fi	Lunga	Popolare	Alta	Preferito	Si	Action
Intermedio	Discreto	Christopher Nolan	Action,Crime,Drama	Lunga	Popolare	Alta	Preferito	Si	Action
Intermedio	Discreto	Christopher Nolan	Adventure,Drama,Sci-Fi	Lunga	Popolare	Alta	Preferito	Si	Adventure
Recente	Discreto	Christopher Nolan	Action,Drama,History	Media	Popolare	Media	Preferito	No	Action
Intermedio	Buono	Christopher Nolan	Mystery,Thriller	Media	Popolare	Alta	Preferito	Si	Mystery
Recente	Discreto	Christopher Nolan	Action,Sci-Fi,Thriller	Lunga	Popolare	Media	Preferito	No	Action

Figure 7: Dataset after Feature Engineering

Le variabili inoltre per essere prese in input dal modello sono state trasformate grazie alla libreria Scikit-learn con il OneHotEncoder. Il *One Hot Encoding* rappresenta una tecnica utilizzata poiché molti modelli di classificazione richiedono input numerici. Questa tecnica consente di convertire le variabili categoriche in array composti esclusivamente da valori 0 e 1.
Ad esempio, consideriamo i valori della feature 'AvgRating' insieme alle eventuali rispettive codifiche:
Alta: [1, 0, 0] Media: [0, 1, 0] Bassa: [0, 0, 1]

3.4 Data Modelling

La variabile indipendente da predire è 'Rating', quindi è necessario formulare il problema come un problema di classificazione.
Dato che il dataset non è particolarmente ampio, si è deciso di addestrare il modello utilizzando il 90% dei dati per l'addestramento (training set) e il restante 10% come test per valutare le prestazioni del modello (test set).
La scelta dell'algoritmo per l'addestramento è stata guidata dalle metriche di valutazione dei modelli testati, le quali sono visibili nella sezione sottostante.

- Albero decisionale

```
Accuracy: 0.46875
Classification Report:
              precision    recall  f1-score   support

      Buono       0.29      0.29      0.29         7
   Discreto       0.36      0.40      0.38        10
     Ottimo       0.64      0.60      0.62        15
```

- Random Forest

```
Accuracy: 0.53125
Classification Report:
              precision    recall  f1-score   support

      Buono       0.00      0.00      0.00         7
   Discreto       0.46      0.60      0.52        10
     Ottimo       0.69      0.73      0.71        15
```

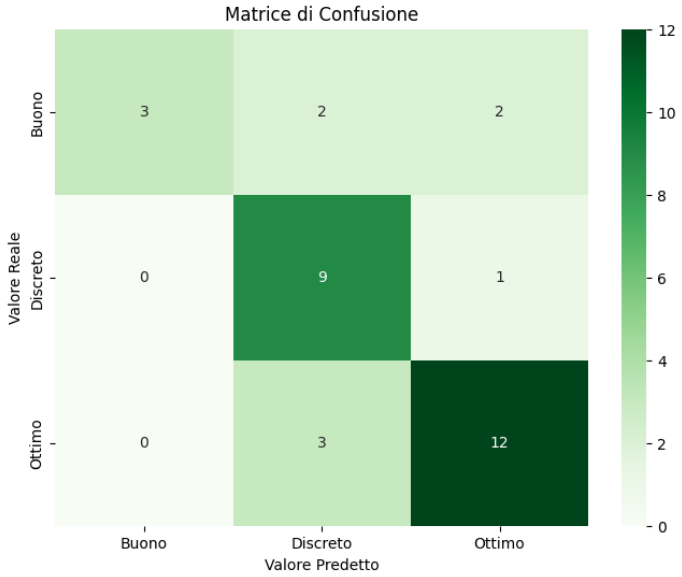
- Support Vector Machine

```
Accuracy: 0.75
Classification Report:
              precision    recall  f1-score   support

      Buono       1.00      0.43      0.60         7
   Discreto       0.64      0.90      0.75        10
     Ottimo       0.80      0.80      0.80        15
```

- k-Nearest Neighbors

```
Accuracy: 0.53125
Classification Report:
              precision    recall  f1-score   support

      Buono       0.20      0.14      0.17         7
   Discreto       0.46      0.60      0.52        10
     Ottimo       0.71      0.67      0.69        15
```

I dati indicano che il modello con l'accuracy più elevata e un equilibrio notevole tra le metriche di valutazione è il **Support Vector Machine**. Per valutare l'efficacia delle predizioni, esaminiamo la matrice di confusione relativa al Support Vector Machine.



Per migliorare le metriche di valutazione del modello, sono state eseguite ulteriori operazioni sui dati. Durante la fase di addestramento, è emerso che il modello raggiunge una maggiore accuratezza senza la presenza della feature 'Apprezzato', poiché si riscontrava un problema di overfitting. Di conseguenza, si è tornati alla fase di data preparation per rimuovere questa feature e successivamente si è proceduto di nuovo con l'addestramento del modello.

3.5 Evaluation

3.5.1 Criticità

Il modello è stato addestrato su un dataset personale, il che significa che sarà in grado di effettuare predizioni solo per l'utente associato a tale dataset. Pertanto, non è utilizzabile da altri utenti con storie diverse.
Nel corso del tempo, l'utente potrebbe aggiungere nuove valutazioni, ma il modello continuerà a basarsi sui dati passati. Le possibili soluzioni a questo problema includono la periodica riaddestramento del modello o l'addestramento continuo per mantenere il modello aggiornato nel tempo.
Va notato che le metriche di valutazione per l'algoritmo potrebbero non essere del tutto ottimali. Tuttavia, è importante considerare che la variabile indipendente nel dataset non segue pattern specifici. La soggettività delle valutazioni dei film può causare cambiamenti arbitrari nelle valutazioni senza un motivo specifico, il che rende la previsione accurata più complessa.

3.5.2 Sviluppi futuri

Si potrebbe considerare la modifica del modello in modo che diventi utilizzabile da tutti gli utenti. In particolare, sarebbe necessario prestare attenzione al bilanciamento dei dati, gestendo situazioni con classi popolose attraverso operazioni di undersampling o oversampling. In prospettiva futura, potrebbe essere vantaggioso introdurre una nuova feature, quale l'attore principale, che potenzialmente potrebbe avere un buon valore predittivo. Va notato che attualmente tale feature non è presente nel dataset attuale.

4 Conclusioni

In sintesi, posso affermare che la realizzazione di questo progetto si è rivelata un'esperienza gratificante e istruttiva. Affrontando la sfida di esplorare nuovi concetti nel campo dell'apprendimento, ho affinato le mie competenze, migliorando non solo nella padronanza del linguaggio Python, ma anche nell'applicazione di librerie specializzate e nell'approccio ingengneristico al problema.